Collective Activity Detection using Hinge-loss Markov Random Fields

Ben London, Sameh Khamis, Stephen H. Bach, Bert Huang, Lise Getoor, Larry Davis
Motivation
Motivation

- Classify the individual actions
Motivation

- Classify the individual actions
- Track the multiple targets
Intuition

- Action transitions are unlikely
Intuition

- Action transitions are typically not arbitrary
Intuition

• Individual actions are consistent in proximity
Intuition

- Individual actions are consistent in proximity
Related Work

- Original action recognition work focused on the isolated person case

  Shuldt et al., ICPR 2004

  Blank et al., CVPR 2005

- Following work investigated either pairwise interactions or group activity as the activity of the majority

  Ryoo and Agarwal, ICCV 2009

  Lan et al., NIPS 2010
Related Work

- More recent work looked at coupling activity recognition, tracking, and scene labeling.

  Choi and Savarese, ECCV 2012

- While others modeled activities at multiple levels: individual, group, and inter-group.

  Amer et al., ECCV 2012

Khamis et al., ECCV 2012
Our Approach

An Introduction to Hinge-loss MRFs and PSL
Our Approach

- Problem needs **scalable** solution that handles complex dependencies and tracking constraints.

- *Hinge-loss Markov Random fields* (HL-MRFs) are a new class of models that meet these goals:
  - Log-concave densities over continuous variables
  - Support fast inference of global solutions
  - New paper on structured prediction at UAI 2013

- *Probabilistic soft logic* (PSL) allows easy encoding of intuitions:
  - Converts logical rules to HL-MRFs
Hinge-loss Markov Random Fields

\[ p(Y|X) = \frac{1}{Z} \exp \left[ -\sum_{j=1}^{m} w_j \max\{\ell_j(Y, X), 0\}^{p_j} \right] \]

- Continuous variables in [0,1]
- Potentials are hinge-loss functions
- Subject to arbitrary linear constraints
- Log-concave!
Inferring Most Probable Explanations

- Objective:

\[
\arg\max_Y p(Y|X) = \arg\min_Y \sum_{j=1}^{m} w_j \max\{ \ell_j(Y, X), 0 \}^{p_j}
\]

- Convex optimization
- Decomposition-based inference algorithm using the ADMM framework
Alternating Direction Method of Multipliers

• Inference with ADMM is fast, scalable, and straightforward

• Optimize subproblems (ground rules) independently, in parallel

• Auxiliary variables enforce consensus across subproblems
Weight Learning

• Various methods to learn from training data:
  o approximate maximum likelihood
  o maximum pseudolikelihood
  o large-margin estimation
  o [Broecheler et al., UAI 2010; Bach et al., UAI 2013]

• State-of-the-art learning performance on
  o Collective classification
  o Social-trust prediction
  o Preference prediction
  o Image reconstruction

• Here we use approximate maximum likelihood
Probabilistic Soft Logic

• HL-MRFs are easy to define
• Hinge-losses can generalize logical operators

1.8: Doing(X, walking) ← SamePerson(X, Y) ∧ Doing(Y, walking)

• Lukasiewicz T-norm
  o \( A \lor B = \min\{1, A + B\} \)
  o \( A \land B = \max\{0, A + B - 1\} \)
Grounding to HL-MRFs

- Ground out first-order rules
  - Variables: soft-truth values of atoms
  - Hinge-loss potentials: weighted *distances to satisfaction* of ground rules

- \[ w : A \rightarrow B \]
  - \[ w : \neg A \lor B \]
  - \[ w \times (1 - \min\{1 - A + B, 1\}) \]
  - \[ w \times \max\{A - B, 0\} \]

- The effect is assignments that satisfy weighted rules more are more probable
A PSL Model for Collective Activity Detection

A Collective Activity Detection Model in PSL
Features: Low-Level

- Histogram of Oriented Gradients (HOG) [Dalal & Triggs, CVPR 2005]

- Describe image patches by a distribution of gradient magnitudes binned by angle

- We train SVMs to predict on HOG features
Features: Low-Level

- Action Context Descriptor (ACD) [Lan et al, NIPS 2010]

- Model context by aggregating SVM outputs on HOG features across multiple spatiotemporal neighborhoods

- E.g., actions like talking cannot be represented by the HOG features of one person
Local Information

- Use low-level detectors

$$w_{\text{local},a}: \text{Doing}(X, a) \leftarrow \text{Detector}(X, a)$$

- E.g.,

$$w_{\text{local},\text{walking}}: \text{Doing}(X, \text{walking}) \leftarrow \text{Detector}(X, \text{walking})$$

$$w_{\text{local},\text{talking}}: \text{Doing}(X, \text{talking}) \leftarrow \text{Detector}(X, \text{talking})$$

$$w_{\text{local},\text{waiting}}: \text{Doing}(X, \text{waiting}) \leftarrow \text{Detector}(X, \text{waiting})$$

- (defined for all actions)
Frame Consistency

• Most people in frame do the same action
• Ground truth is aggregate of descriptors

\[ w_{\text{frame},a} : \text{Doing}(X, a) \leftarrow \text{Frame}(X, F) \land \text{FrameAction}(F, a) \]
Effect of Proximity

- People that are close (in frame) are likely doing the same action

\[ w_{\text{prox},a} : \text{Doing}(X, a) \leftarrow \text{Close}(X, Y) \land \text{Doing}(Y, a) \]

- Closeness is measured via a radial basis function
Tracking

• Persistence rules
  o People are likely to continue doing the same action

  \[ w_{\text{persist}} : \text{Doing}(Y, a) \leftarrow \text{SamePerson}(X, Y) \land \text{Doing}(X, a) \]

  o Requires identity maintenance for SamePerson

• Identity maintenance

  \[ w_{\text{id}} : \text{Same}(X, Y) \leftarrow \text{Sequential}(X, Y) \land \text{Close}(X, Y) \]
Action Transitions

• Can define rules for transitioning between actions

\[ w_{\text{trans}, a, b} : \text{Doing}(Y, b) \leftarrow \text{SamePerson}(X, Y) \land \text{Doing}(X, a) \]

• Defined over all pairs of actions (a,b)
• Effect is similar to the state transition matrix of an HMM
Priors and Constraints

• Prior beliefs
  o Encode prior beliefs about SamePerson and Doing predicates

\[ w : \neg \text{SamePerson}(X, Y) \quad w : \neg \text{Doing}(X, a) \]

• Constraints
  o Functional constraint on Doing ensures that soft-truth values for each person sum to 1
  o Partial-functional constraint on SamePerson ensures that soft-truth values for each person sum to at most 1
Experiments
Dataset

• University of Michigan, “Collective Activity”
• Annotated activities, poses, trajectories
  o We don’t use poses, trajectories
  o We only use activity annotations for training
• 2 common splits:
  o 5-label: [ crossing, walking, waiting, talking, queueing ]
    • 44 sequences
  o 6-label: [ crossing, waiting, talking, queueing, dancing, jogging ]
    • 63 sequences

http://www.eecs.umich.edu/vision/activity-dataset.html
PSL Model

\[ w_{id} : \text{Same}(X, Y) \leftarrow \text{Sequential}(X, Y) \land \text{Close}(X, Y) \]

\[ w_{idprior} : \sim\text{SamePerson}(X, Y) \]

For all actions \(a\):

\[ w_{\text{local},a} : \text{Doing}(X, a) \leftarrow \text{Detector}(X, a) \]

\[ w_{\text{frame},a} : \text{Doing}(X, a) \leftarrow \text{Frame}(X, F) \land \text{FrameAction}(F, a) \]

\[ w_{\text{prox},a} : \text{Doing}(X, a) \leftarrow \text{Close}(X, Y) \land \text{Doing}(Y, a) \]

\[ w_{\text{persist},a} : \text{Doing}(Y, a) \leftarrow \text{SamePerson}(X, Y) \land \text{Doing}(X, a) \]

\[ w_{\text{prior},a} : \sim\text{Doing}(X, a) \]
Methodology

• Measure benefit of high-level reasoning
  o One model using HOG SVM scores, another using ACD SVM scores
  o Measure lift over low-level detectors

• Leave-one-out cross-validation
  o Train on all but one sequence
  o Test on hold-out
  o Accumulate test statistics over all hold-outs
    • Compensates for varying lengths and label distributions
## Results

<table>
<thead>
<tr>
<th></th>
<th>5-Action</th>
<th>6-Action</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>F1</td>
</tr>
<tr>
<td>HOG SVM</td>
<td>0.474</td>
<td>0.481</td>
</tr>
<tr>
<td>HL-MRF + HOG</td>
<td>0.598</td>
<td>0.603</td>
</tr>
<tr>
<td>ACD SVM</td>
<td>0.675</td>
<td>0.678</td>
</tr>
<tr>
<td>HL-MRF + ACD</td>
<td>0.692</td>
<td>0.693</td>
</tr>
</tbody>
</table>
What about MLNs?

• Also compare against an identical Markov logic network (MLN) model
  o Inference and MLE in MLNs are generally intractable
  o MaxWalkSat for learning
  o MCSAT for test-time inference
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>5-Action Accuracy</th>
<th>5-Action F1</th>
<th>6-Action Accuracy</th>
<th>6-Action F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG SVM</td>
<td>0.474</td>
<td>0.481</td>
<td>0.596</td>
<td>0.582</td>
</tr>
<tr>
<td>MLN + HOG</td>
<td>0.657</td>
<td>0.657</td>
<td>0.809</td>
<td>0.803</td>
</tr>
<tr>
<td>HL-MRF + HOG</td>
<td>0.598</td>
<td>0.603</td>
<td>0.793</td>
<td>0.789</td>
</tr>
<tr>
<td>ACD SVM</td>
<td>0.675</td>
<td>0.678</td>
<td>0.835</td>
<td>0.835</td>
</tr>
<tr>
<td>MLN + ACD</td>
<td>0.687</td>
<td>0.685</td>
<td>0.850</td>
<td>0.850</td>
</tr>
<tr>
<td>HL-MRF + ACD</td>
<td><strong>0.692</strong></td>
<td><strong>0.693</strong></td>
<td><strong>0.860</strong></td>
<td><strong>0.860</strong></td>
</tr>
</tbody>
</table>
Speed

Average running time

<table>
<thead>
<tr>
<th></th>
<th>Cora</th>
<th>Citeseer</th>
<th>Epinions</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLN</td>
<td>110.9 s</td>
<td>184.3 s</td>
<td>212.4 s</td>
<td>344.2 s</td>
</tr>
<tr>
<td>HL-MRF</td>
<td>0.4 s</td>
<td>0.7 s</td>
<td>1.2 s</td>
<td>0.6 s</td>
</tr>
</tbody>
</table>

[Bach et al., UAI 2013]

- MLN inference is **slow**
  - MCSAT is poly-time, but slow

- HL-MRF inference is **fast**
  - In practice, we find that inference scales linearly with the number of potentials
Improved PSL Model

• Scene consistency
  o Certain sequences tend to have a single majority action
  o Improved performance in [Khamis et al., ECCV 2012]

• In-frame/sequence interactions
  o E.g., Maybe walking and crossing frequently co-occur together?

• Latent variables
  o E.g., Group actors into same-action clusters, reason about cluster interactions
Conclusion

• HL-MRFs are a powerful class of graphical models
  o Capable of fast MPE inference
  o Faster inference than discrete models (e.g., MLNs)

• PSL facilitates easy construction of HL-MRFs
  o First-order-logic syntax

• Using HL-MRFs/PSL for high-level vision yields significant improvement over low/mid-level detectors
Thank you!


- S. Khamis, V. I. Morariu, L. S. Davis. Combining Per-Frame and Per-Track Cues for Multi-Person Action Recognition. ECCV, 2012
- T. Lan, Y. Wang, W. Yang, G. Mori: Beyond Actions: Discriminative Models for Contextual Group Activities. NIPS 2010
- C. Schult, I. Laptev, B. Caputo. Recognizing Human Actions: A Local SVM Approach. ICPR, 2004